PATENT APPLICATION
DOCKET NO.: 1285-0124US
ALC-139137

"EXPRESS MAIL" Mailing Label No..EV331251355US.........
Date of Deposit...........AUGUST 21, 2003..............

MULTI-TIME SCALE ADAPTIVE INTERNET PROTOCOL ROUTING SYSTEM AND METHOD

BACKGROUND OF THE INVENTION

Technical Field of the Invention

[0001] The present invention generally relates to internet protocol ("IP") routers. More particularly, and not by way of any limitation, the present invention is directed to an architecture for implementing multi-time scale adaptive IP routers.

Description of Related Art

[0002] The deployment of complex value-added services on the Internet requires support for appropriate time-scale resource management and control on an on-going basis after the initial setup time of the network. Frequent interaction between services and the network calls for identification and implementation of sophisticated mechanisms that can allocate resources (e.g., links, switch capacity, etc.) optimally, isolate or dynamically share resources in a controlled fashion among various users, and change the use and allocation of resources in response to changes in conditions of the network and user requirements over time.

[0003] For QoS adaptation, the most developed adaptive resource allocation in IP is based on end-to-end feedback systems, such as TCP and ATM congestion control. In end-to-end feedback systems, the receiver must inform the sender that there is a problem when it arises. The main drawback to such systems is that often by the time adaptation has occurred, the condition might already have changed. Accordingly, such systems are typically not sufficiently fast.

[0004] Adaptation of the switching node behavior on the basis of DiffServ, Per-Hop Behaviors ("PHBs"), and variants of Random Early Detection ("RED") are commercially available. With these, policies and priorities are associated with the packets. When congestion occurs, packets are selected for dropping based on these policies and priorities. However, these are neither measurement-based nor are they designed to work in a coordinated fashion at multi-time scale.

[0005] IEEE has proposed a standard (IEEE 1520) that brings layering of IP router functionality for programmability and adaptation. This standard defines an overall infrastructure, but not a method or system for performing adaptation. The ForCES working group in IETF has proposed ideas that separate the control elements from forwarding elements in a way that can help introduce dynamic behavior at the node level. Also, network equipment manufacturers are currently beginning to support measurement-based adaptation. However, none of the foregoing solutions support implementation of a multi-time scale adaptive router.

SUMMARY OF THE INVENTION

[0006] Accordingly, the present invention advantageously provides an architecture for implementing multi-time scale adaptive IP routers. One embodiment is a packet router that supports multi-time scale resource management. The packet router comprises a management agent that manages differentiated services policy information database operable to store policies on forwarding packets in the packet router; a resource server system that controls forwarding of packets in the packet router based on adaptive selections of policies from the policy information database; a flow measurement system that monitors packet flows through the packet router and generates statistics reports which affect the resource server systems selection of control; and a hardware forwarding engine that receives and forwards packets in response to the resource server system controls.

[0007] Another embodiment is a system for supporting multi-time scale resource management in a packet router. The system comprises means for managing a differentiated services policy information database that stores policies on forwarding packets in the packet router; means for controlling forwarding of packets in the packet router based on adaptive selections of policies from the policy information database; means for monitoring packet flows through the packet router; means for generating statistic reports that affect the resource server systems selection of control; and means for receiving and forwarding packets in response to the resource server system controls.

[0008] Another embodiment is a method of providing multi-time scale resource management in a packet router. The method comprises managing a differentiated services policy information database that stores policies on forwarding packets in the packet router; controlling forwarding of packets in the packet router based on adaptive selections of policies from the policy information database; monitoring packet flows through the packet router; generating statistic reports that affect the forwarding of packets in the packet router; and receiving and forwarding packets in response to the forwarding of packets in the packet router.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] A more complete understanding of the present invention may be had by reference to the following Detailed Description when taken in conjunction with the accompanying drawings wherein:

[0010] FIG. 1A is a block diagram of one embodiment of an enhanced IP router for supporting multi-times scale resource management in an IP network;

[0011] FIG. 1B is a block diagram of one embodiment of an IP network in which the router of FIG. 1A may be implemented;

[0012] FIG. 2 is a block diagram of one embodiment of a Resource Server System of the router of FIG. 1A;

[0013] FIG. 3A is a block diagram of one embodiment of a Flow Measurement System of the router of FIG. 1A;

[0014] FIG. 3B is illustrates the functions of IP Flow Monitor and Measuring group modules of the router of FIG. 1A;

[0015] FIG. 4 is a block diagram of one embodiment of a Monitor Data Collector ("MDC") of the router of FIG. 1A; and

[0016] FIG. 5 is a flowchart of one example of the operation of the router of FIG. 1A.

DETAILED DESCRIPTION OF THE DRAWINGS

[0017] In the drawings, like or similar elements are designated with identical reference numerals throughout the several views thereof, and the various elements depicted are not necessarily drawn to scale.

[0018] FIG. 1A illustrates a framework for an enhanced IP switching node, or router, 100 that supports multi-time scale resource management in an IP network. For purposes of illustration, operation of the node 100 will be described in the case of DiffServ-based quality of service ("QoS") adaptation through Policy Information Base ("PIB") and Service Level Agreement ("SLA") management functionality. As illustrated in FIG. 1A, the node 100 includes a framework for traffic monitors and mechanisms for closed loop feedback at data, flow, and network time-scale level controls. The node 100 comprises four primary components, including a management agent ("MA") 102, which resides in a management plane 103; a resource server system ("RSS") 104 and a flow measurement system ("FMS") 106, both of which reside in a control plane 107; and hardware forwarding units ("HFUs") 108, which reside in a data plane 109. The configuration of the node 100 gives rise to static and dynamic functionality portions of the node 100, as will be described in greater detail with reference to FIG. 2.

[0019] In the case of DiffServ QoS, the MA 102 performs DiffServ PIB management functionality. The RSS 104 on the control plane 107 controls IP packets through the data plane 109 based on a PIB table. The FMS 106 monitors packet flows through the HFUs 108 on the data plane 109. The elements 102, 104, 106, and 108, will be described in greater detail below with reference to FIGs. 2, 3, and 4.

[0020] FIG. 1B illustrates an IP network 150 comprising a plurality of adaptive core routers 152 and adaptive edge routers 154 functioning under the control of a single central server ("CS") 156. The CS 156 comprises provides a network-wide management and provisioning entity, such as a bandwidth broker. It will be recognized that, in accordance with one embodiment, the architecture of each of the routers 152, 154, is as illustrated in FIG. 1A and described in greater detail below. In general, an "adaptive router" is equivalent to a standard router with the addition of a local monitor and a local controller, as also described below.

[0021] The routers 152, 154, are interconnected via links 158. Additionally, the edge routers 154 may be connected to end systems 160. The core routers 152 support data level adaptations, the edge routers 154 support flow level adaptations, and the CS 156 supports network level adaptations.

[0022] FIG. 2 illustrates the architecture of the RSS 104. As shown in FIG. 2, the RSS 104 includes a service component ("SC") 200 with three distinct layers, including a Service Abstraction Layer ("SAL") 202, a Resource Abstraction Library Layer ("RALL") 204, and a Logical Hardware Abstraction Layer ("LHAL") 206, and a Hardware Interface Control Component ("HICC") 208. Again, in the context of DiffServ QoS, the SAL 202 of the SC 200 uses PIB tables (not shown) to define a configuration and management table (not shown) of the DiffServ interface in terms of a Traffic Control Block ("TCB") and selects a sequence of resource abstraction elements. The SAL 202 is also controlled by a Dynamic

Component ("DC") 210 to adapt to dynamic service requirements and resource conditions through actions such as adjusting bandwidth, loop gain, and sampling rate without generating instability.

[0023] The RALL 204 of the SC 200 has an open architecture with a library of router and resource management algorithms (e.g., classifier, meter, queue) selected by the TCB. The RALL 204 provides control to a Hardware Forwarding Manager ("HFM") 212 of the HICC 208 via a Logical Hardware Abstraction Layer ("LHAL") 214. The RALL 204 contains only the algorithms that have no contents of the physical layer hardware. The LHAL 206 contains only the logical hardware elements (i.e., the finest units to compose the network element resources and service function behaviors), which are monitored by the MDC (FIG. 3A) of the DC 210. The HICC 208, in turn, contains the HCU 214, which is connected through an interface driver 216 to the RALL 204, and the HFM 212, which is connected through an interface driver 218 to the LHAL 206 and the DC 210. The HCU 214 of the HICC 208 controls the HFU 108 packets via the HFM 212. The RALL 204 of the SC 200 can bypass the HCU 214 and directly control the HFM 212 via the LHAL 206. Generally, the vendor's line cards with interface driver software will support the HICC 208. The advantages of such an architecture for the RSS 104 are described in detail below.

[0024] FIG. 3A illustrates the architecture of the FMS 106. The FMS 106 contains the DC 210, an FMS Reports Buffer ("FMSRB") 300 and a Policy Provision Info Buffer ("PPIB") 302. The DC 210 in turn is comprised of three layers, including a Monitor Resource Controller ("MRC") 304, a Monitor Resource Abstraction Library ("MRAL") 306, and a Monitor Data Collector/Data Source Controller ("MDC") 308. In the DiffServ adaptation, the DC 210 receives interpreted service level agreement ("SLA") from the PPIB 302 and sends flow measurement statistics ("FMS") reports/alarms to the FMSRB 304.

[0025] The MRC 304 receives interpreted SLA from the PPIB 302 and produces Observation Domain Blocks (ODBs) (FIG. 4) in terms of a sequence of MRAL 306 elements (i.e. Metering Process, Selection Criteria for Flow Export, Export Process, and Data Collector). By executing ODB sequence elements, an FMS system is performed at line speed, broader bandwidth with open software architecture and open hardware architecture.

[0026] The MRAL 306 acts as a real-time monitor executive to collect and aggregate data sources ("DS") statistics, prepare DS statistics reports, distribute protocols, relocate hardware resources, adjust an AQoS scaler/loop, and determine data collection sampling rates. The MRAL 306 also controls the HFM 212 of the RSS 104 via the Interface Driver 218. A connection loop from the MRAL 306 to the HFM 212, from the HFM 212 to the

HFUs 108, from the HFUs 108 to the MDC 308 and from the MDC 308 back to the MRAL 306 forms a dynamic adaptive scalable closed loop, which is based on FMS SLA and is dynamically controlled by real-time traffic flows. The MRAL 306 is an open architecture design; it allows FMS algorithm insertion/deletion to the system without disturbing system software or hardware performances. The MRAL 306 includes a DS Monitor SLA Interpretation and Statistics Reporting group, a Counter Aggregation control group, a Capabilities group, a Computer Resource group, an AQoS Scaler/Loop Gain Adjuster group, a Data Collector group, and an IP Flow and Measuring group.

[0027] The DS Monitor SLA Interpretation and Statistics Reporting group communicates with a Policy Provision Agent and performs SLA interpretation to modify a Service Component Policy Information Base (PIB) table based on the dynamic resource management of the Capabilities and Computing Resource groups in order to meet the SLA requirements. It also reports the DS statistics, Protocol statistics, Network Protocol Host statistics, and Application Protocol Matrix statistics by invoking the Counter Aggregation Control and Data Collection groups.

[0028] The Counter Aggregation control group controls how individual DiffServ codepoint counters are aggregated in data collections. The Capabilities group describes the groups that are supported by the agent on at least one data source and includes the following modules: (1)

total capability calculation; (2) effective bandwidth determination; and (3) minimized cost based on traffic weight class.

[0029] The Computer Resource group comprises the computer or hardware resources that can be dynamically allocated to meet the SLA requirements through invoking the Capabilities group module. The AQoS Scaler/Loop Gain Adjuster group is based on invoking of IP Flow Monitor & Measuring group module to determine the sampling rate, bandwidth, loop gain, and stability. It is a real-time adaptive closed loop control, which is based on IP DiffServ SLA, IP flow measurement requirements (from measurement SLA), and observation points.

[0030] The Data Collector group controls how individual statistical collections are maintained by the agent and reported to management applications. The Data Collection group collects DS statistics, Protocol statistics, Network Protocol Host statistics and Application Protocol Matrix statistics. The IP Flow Monitor & Measuring group provides a standard way of exporting information related to IP flows for monitoring and measuring and includes the following modules: (1) Flow Classification; (2) Packets Selection Criteria: (3) Metering Process; (4) Selection Criteria for Flow Export; (5) Flow Expiration, (6) Packet Counter and Dropper Counter; (7) Time Stamping; (8) Sampling Method; (9) Exporting Process; (10) Collecting Process; and (10) Effective Bandwidth Measurement.

[0031] FIG. 3B illustrates the functions performed by the IP Flow Monitor & Measuring group modules. In general, these modules perform metering (probing) process functions 350, flow export selection criteria functions 352, collector functions 354, application functions 356, IP flow information protocol selection criteria functions 358, an export model function 360, a collector crash detection and recovery function 362, a collector redundancy function 364, and security functions 366.

[0032] FIG. 4 illustrates the architecture of the MDC 308. The MDC 308 is generally deployed into an *MxN* matrix of processor elements ("PEs") based on the number of observation points 400 of the HFUs 108. In the DiffServ implementation, the HFUs 108 on the data plane 109 can be partitioned into the desired number of observation points 400 based on the SLA received from the PPIB 302 processed by the MRC 304 using an IP Flow Monitoring and Measuring Group Module 402 and DS Monitor SLA Statistics Reporting Group Modules 404 to form Observation Domain Blocks ("ODBs") 406(1)-406(*k*). The IP flow MRAL library 407 of the MRAL 306 acts as a real-time monitor executive by performing functions such as collecting, aggregating, generating DS statistics reports, distributing protocols, relocating hardware resources, adjusting QoS levels, and determining data collection sampling rates. The MRAL 306 also controls the HFM 212 of the RSS 104 via an interface driver 218.

[0033] The MRAL 306-HFM 212-HFU 108-MDCL 308-MRAL 306 connection forms a dynamic adaptive scalable closed loop. The IP flow MRAL library 407 is an open architecture design allowing real-time flow measurement algorithm insertion/deletion without disturbing system software or hardware. Again, in the case of DiffServ QoS, the MRC 304 receives interpreted SLA from the PPIB 302 and produces ODBs in terms of a sequence of MRAL 306 elements (i.e., the metering process, section criteria for flow export, export process, data collector, etc.). By executing ODB sequence elements, the FMS 106 performs at line speed and broader bandwidth with open software architecture and open hardware architecture.

[0034] A flowchart illustrating an example of a typical operational scenario that is supported by the adaptive router 100 in an IP network is set forth in FIG. 5. In step 500, the overall resource configuration of the network is be set up by a centralized unit, such as the CS 156 (FIG. 1), and passed onto each router's configuration through the MA 402. In step 502, traffic monitors embedded in the enhanced routers, such as the router 100, measure the real-time traffic loads per a prescribed set of policy rules. In step 504, individual nodes trigger actions on the basis of preset conditions at data, flow, and network (reporting to the centralized unit) time scales. In step 508, the controller in the nodes and the centralized unit will then adjust service rates to match the requirements of the user and the network conditions in a coordinated manner.

[0035] Several advantages over the prior art are realized by the embodiments described herein. Such advantages include that the connection in the MRAL-HFM-HFU-MDU, along with the reporting capability of the node 100 to the centralized controller forms feedback mechanisms for enabling an adaptive control system at multiple time scales. In addition, the layered architecture of the RS and FMS affords the advantages set forth below.

1. The SAL and MRC have zero hardware contents, thus allowing functions such as SLA to be implemented directly without the knowledge of system low level hardware interfaces. This will save time in complying with service contracts.
2. The RALL and MRAL algorithm insertion/deletion has zero effect on the system software, thus speeding up system upgrade or developments.
3. The HICC and HFU are implemented with hardware (e.g., high-speed network processors or FPGAs) for high speed and high performance gains.
4. Automatic translation of SAL to LHAL through the use of RALL Library elements speeds up the software coding process.
5. The LHAL has zero contents of network element resources and service functional behaviors,

thus universal hardware interface to any HICCs and HFUs can be implemented.

6. The HICC and HFU are vendor-supported hardware (line cards) and software (APIs and interface drivers). This provides shorter system hardware and software integration time.
7. Observation points of HFU can be configured into an *MxN* matrix of PEs for achieving parallel and pipelined processing of data collection.

[0036] Based upon the foregoing Detailed Description, it should be readily apparent that the present invention advantageously provides an architecture for implementing multi-time scale adaptive IP routers.

[0037] It is believed that the operation and construction of the present invention will be apparent from the Detailed Description set forth above. While the exemplary embodiments of the invention shown and described have been characterized as being preferred, it should be readily understood that various changes and modifications could be made therein without departing from the scope of the present invention as set forth in the following claims.